IBM-Northwestern@TRECVID 2014: Surveillance Event Detection(SED)

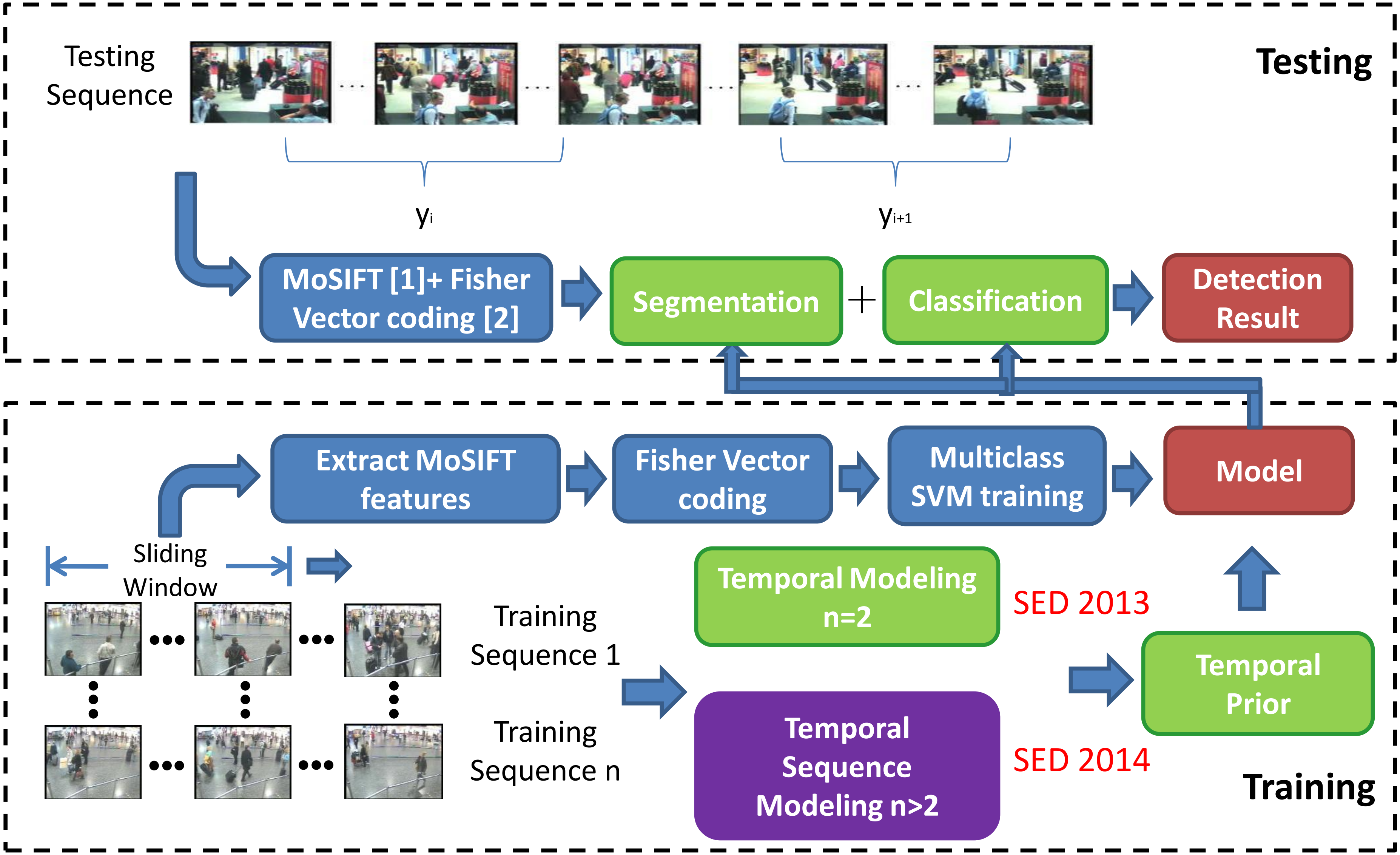
Yu Cheng ɨ*, Jingjing Liu ɨ, Lisa Brown ɨ, Quanfu Fan ɨ, Rogerio Feris ɨ, Alok Choudhary *, Sharath Pankanti ɨ

ɨ IBM Research
*Northwestern University

Outline

- Retrospective Event Detection
  - Sequence Modeling for Event Detection
  - System Overview
  - Performance Evaluation
- Interactive Event Detection
  - Interactive Visualization
  - Risk Ranking
  - Performance Evaluation

System Overview



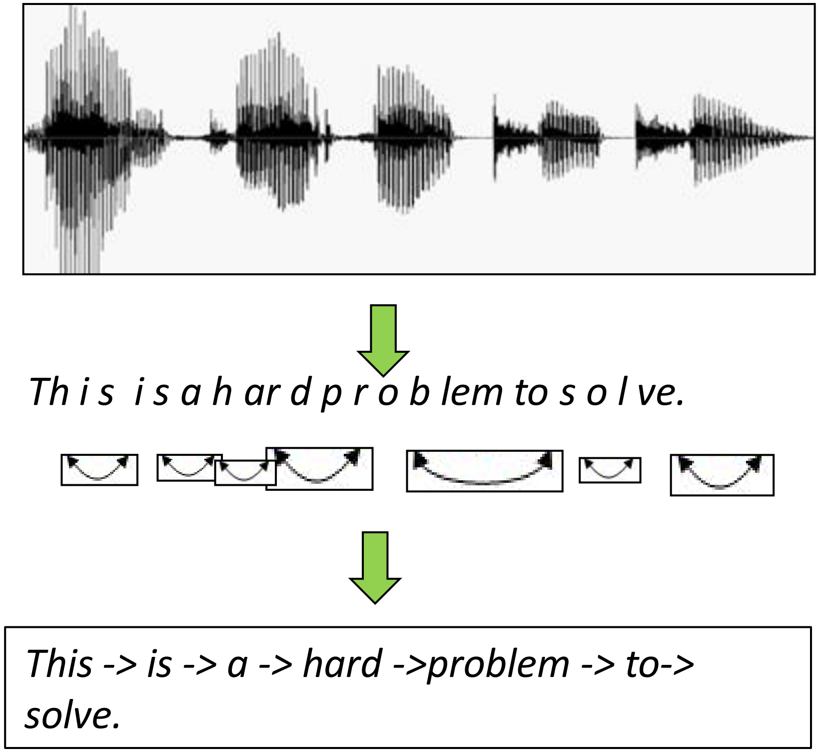
Sequence Temporal Modeling

- Emphasises:
  - Long distance temporal relationship Vs. Short range temporal contexts.
  - Modeling on visual words level Vs. Modeling on event level.

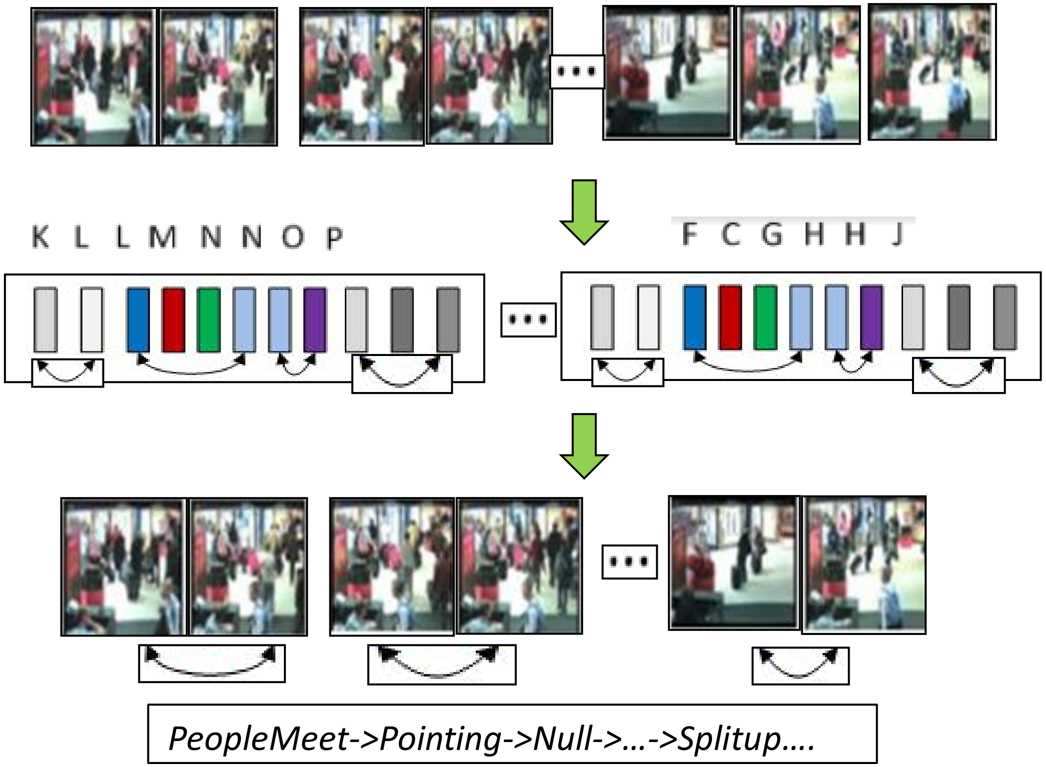
Primary Runs Results	IBM 2014	IBM2013
	ActDCR	ActDCR
CellToEar	0.9914	1.0007
Embrace	0.7456	0.8
ObjectPut	1.0046	1.004
PeopleMeet	0.8160	1.0361
PeopleSplitUp	0.8278	0.8433
PersonRuns	0.8111	0.8346
Pointing	1.0050	1.0175

Motivation

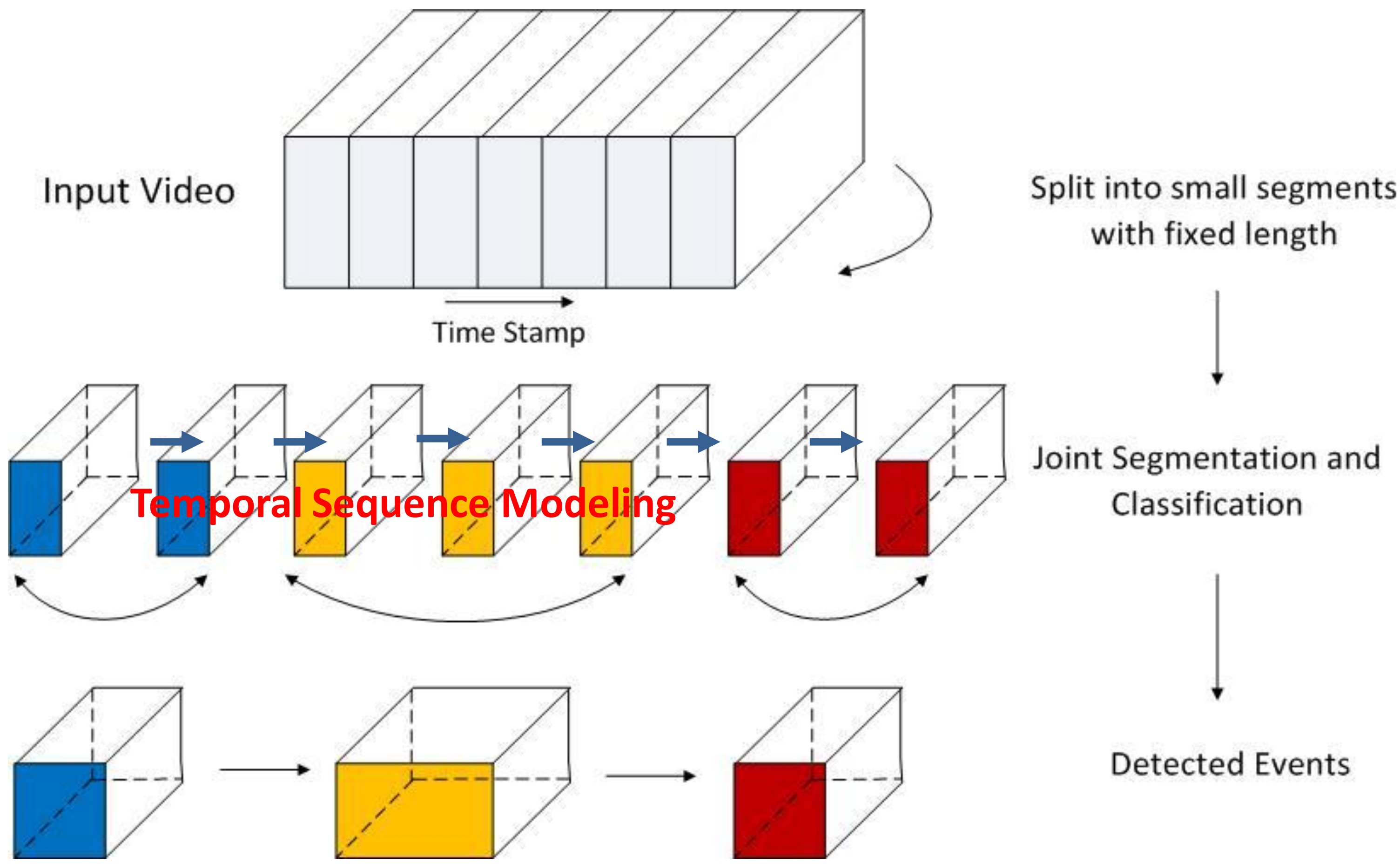
Speech Recognition



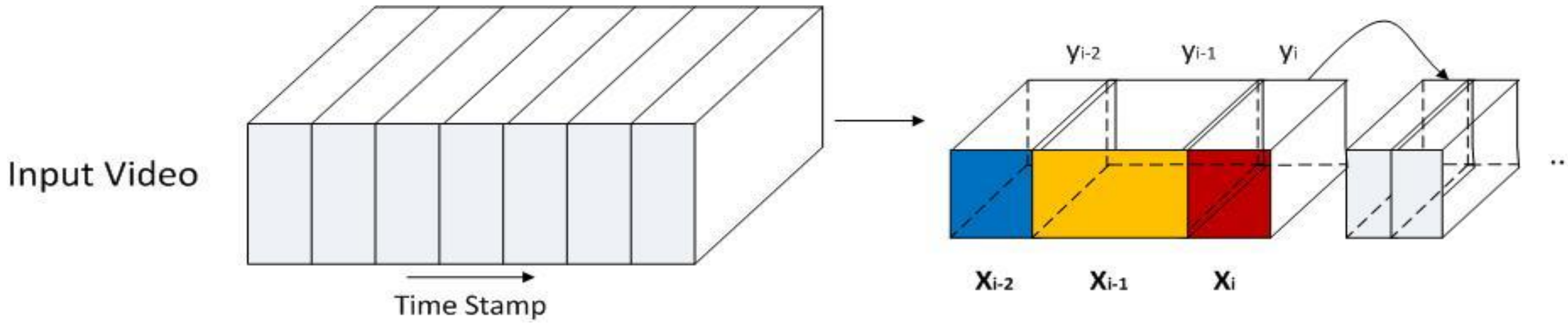
Video Event Detection



Our Method – Framework



Problem Formulation



$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$: detections of video sequence

$\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_m\}$: event class labels of each detection

Joint event classification and segmentation by maximizing

$$f(\mathbf{Y}, \mathbf{X}, \mathbf{Z}) = \sum_{i=1}^{m} \underbrace{\varphi(\mathbf{y_i}|\mathbf{x_i})}_{(1)} + \mu \sum_{1 \le k \le i-1}^{l} \underbrace{p(\mathbf{z_i}|\mathbf{z_{i-k}}, \cdots, \mathbf{z_{i-1}})}_{(2)}$$

$\mathbf{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_l\}$: visual sequence (visual words or events label)

Classification: multi-class SVM

Solver: dynamic programming (*M. Hoai et al, 2011*)

Temporal Sequence Modeling

a) Markov Model

$$P(x_{1:N}) \;=\; \prod_{i=1}^{N} P(x_i|x_1,\ldots x_{i-1}) = P(x_1)P(x_2|x_1)P(x_3|x_2)P(x_4|x_3)\ldots$$

b) Non-Markov Model

$$P(x_{1:N}) \;=\; \prod_{i=1}^{N} P(x_i|x_1,\ldots x_{i-1}) = P(x_1)P(x_2|x_1)P(x_3|x_2,x_1)P(x_4|x_3,\ldots x_1)\ldots$$

Statistical counting in Markov model (i.e. **nth-order** when len(u)=n)

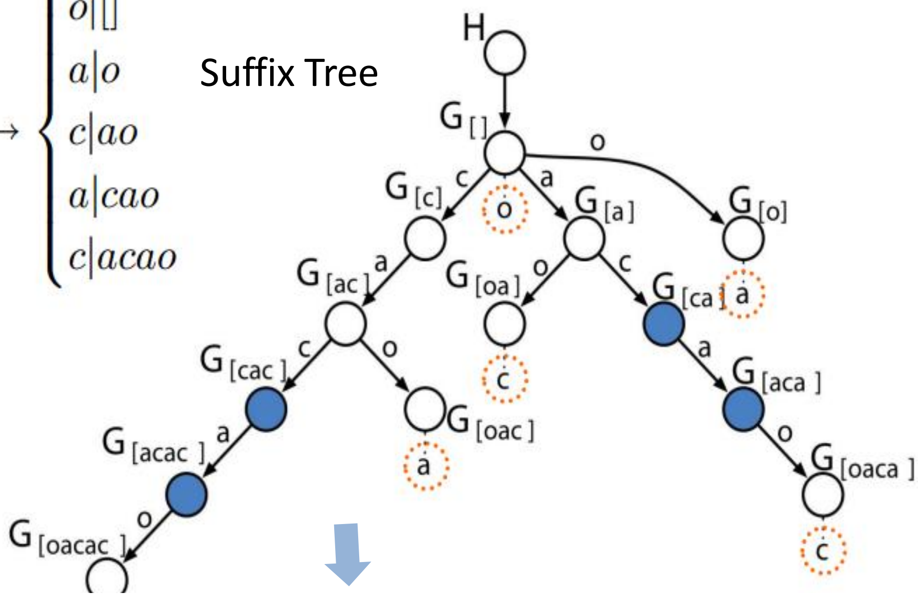
$$G_{\mathbf{u}}(s) = \frac{N(\mathbf{u}s)}{\sum_{s'\in\Sigma} N(\mathbf{u}s')} \qquad \Sigma = x_1, x_2, \ldots, x_T$$

Issues: sparsity, overfitting and scalability

Sequence Memoizer (SM)

$$oacac \rightarrow \begin{cases} o|[] \\ a|o \\ c|ao \\ a|cao \\ c|acao \end{cases}$$

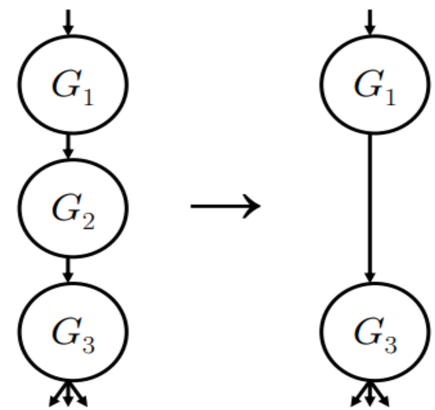
Suffix Tree



$$\begin{aligned} \mathcal{G}_{[]} \mid d_0, \mathcal{U} &\sim \mathrm{PY}(d_0, 0, \mathcal{U}) \\ \mathcal{G}_{[\mathbf{u}]} \mid d_{|\mathbf{u}|}, \mathcal{G}_{[\sigma(\mathbf{u})]} &\sim \mathrm{PY}(d_{|\mathbf{u}|}, 0, \mathcal{G}_{[\sigma(\mathbf{u})]}) \\ x_i \mid \mathbf{x}_{1:i-1} = \mathbf{u} &\sim \mathcal{G}_{[\mathbf{u}]} \\ i = 1, \ldots, T & \\ \forall \mathbf{u} \in \Sigma^{+} & \end{aligned}$$

Hirearchical PYP: G[u] is a PYP with a base of the PYP its parent.

Marginization (efficiency)

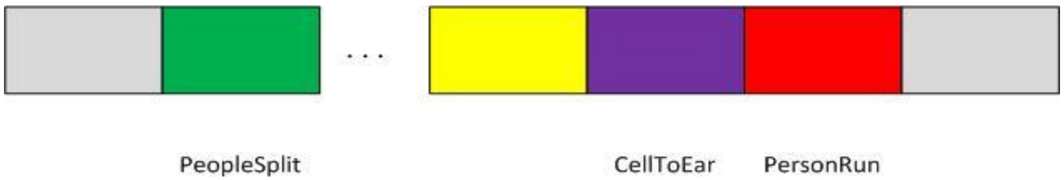


$$G_2|G_1 \sim \mathrm{PY}(d_1, 0, G_1)$$
$$G_3|G_2 \sim \mathrm{PY}(d_2, 0, G_2)$$

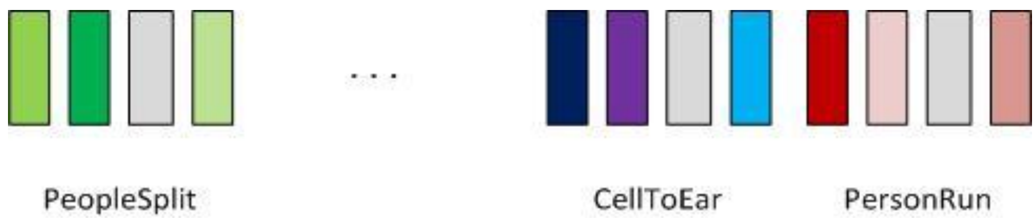
$$G_3|G_1 \sim \mathrm{PY}(d_1 d_2, 0, G_1)$$

(*Frank et al 2009*)

Modeling on event vs. on visual words



$$p(y_i | y_1, \cdots, y_{i-1})$$



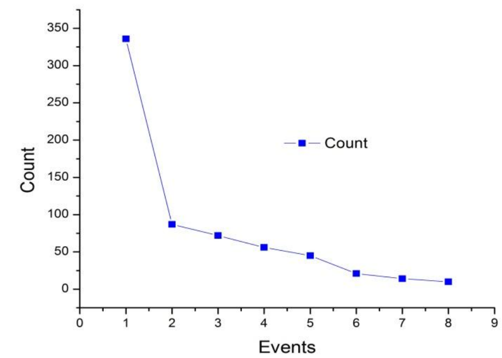
$$p(z_i | z_{i-k} \cdots z_{i-1}) = p(w_{t_i^1}, \ldots, w_{t_i^2} | w_{t_{i-k}^1}, \ldots, w_{t_{i-1}^2})$$

z_i : the *i*-th segmentation

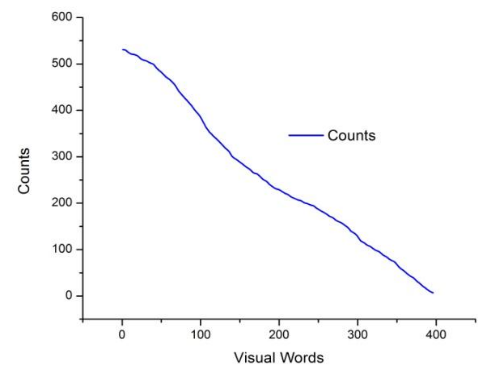
w_i : the *i*-th visual word in z_i

[G. Zipf. Selective studies and the principle of relative frequency in language. 1932.]

Poor Granularity



Good Granularity



Performance Evaluation

Primary Runs Results	IBM 2014		Others' Best 2014	IBM2013
	Ranking	ActDCR	ActDCR	ActDCR
CellToEar	1	0.9914	1.0032	1.0007
Embrace	1	0.7456	0.7845	0.8
ObjectPut	2	1.0046	1.0023	1.004
PeopleMeet	1	0.8160	0.9125	1.0361
PeopleSplitUp	2	0.8278	0.8134	0.8433
PersonRuns	1	0.8111	0.8339	0.8346
Pointing	2	1.0050	1.0040	1.0175

- Compared to our last year's system (IBM 2013):
  - this year system got improvement over 6/7 events (actual DCR of primary run).
- Compared to this year other teams' results (Others' Best 2014):
  - our system leads in 4/7 events (actual DCR of primary run).

Outline

- Retrospective Event Detection
  - System Overview
  - Temporal Modeling for Event Detection
  - Performance Evaluation
- Interactive Event Detection
  - Interactive Visualization System
  - Risk Ranking
  - Performance Evaluation

Detection Results Visualization

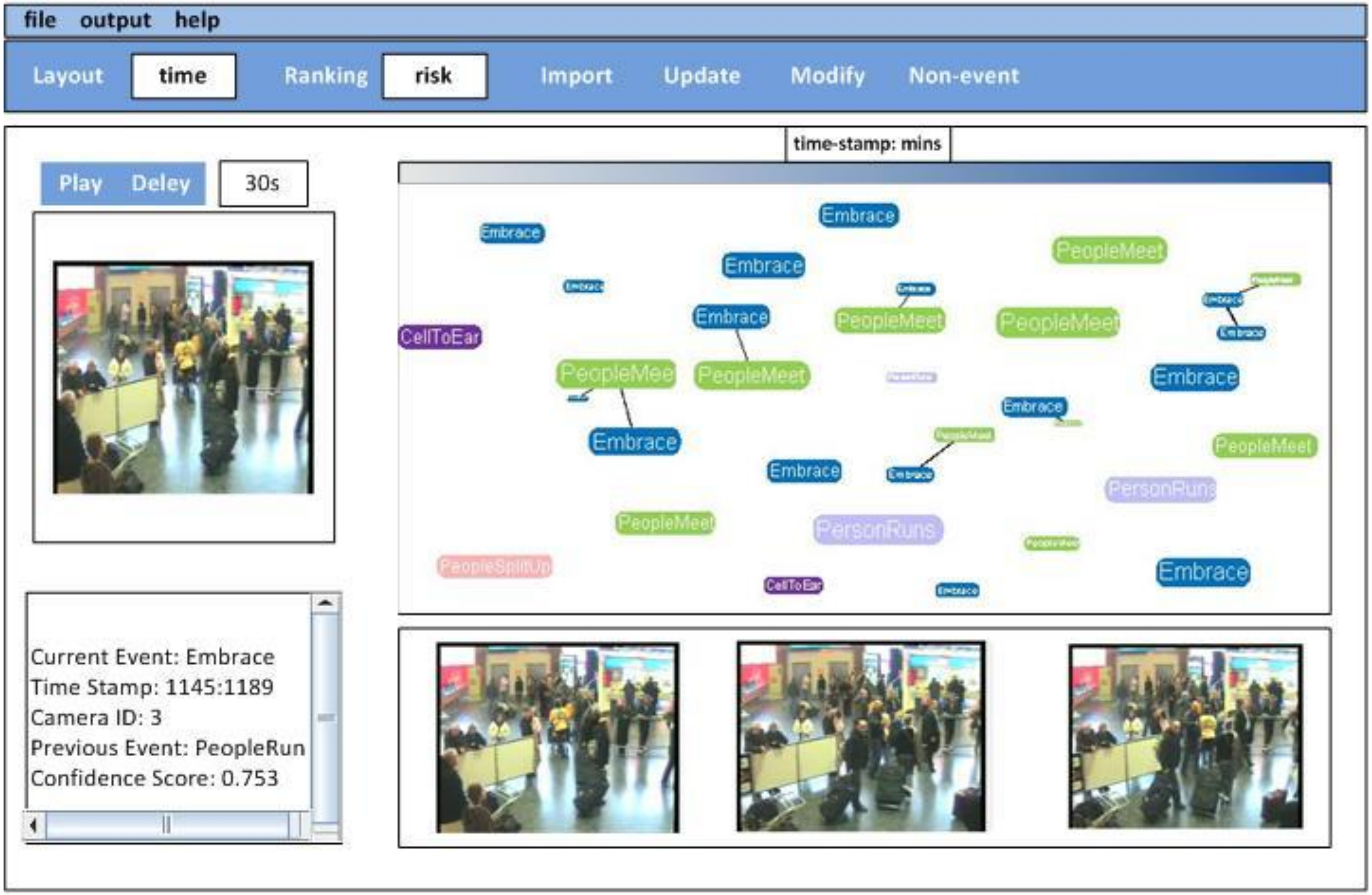
- Motivation:
  - Instead of looking at a single event alone, how can we represent events with strong temporal patterns?
    - E.g. two detected events "Peoplemeet" and "pointing" may exist successively, if we look at them together, it will be effective and efficient.
  - Given thousands of events, how can we differentiate them and present more informative ones to users?
    - E.g. correct some wrong events will get more credit from DCR score, for example, "embrace" → "peoplemeet" vs. "pointing" → "nonevent".

Multiple Detections Visualization

- Objective:
  - To find visualization methods that enable multiple events representation.

- Solution:
  - Visualize the events in a graph-based layout: each node is an individual event and the edge between them representing the temporal relation.

Event-specific Detection Visualization



Visualization with Temporal Relation

People Meet

Pointing

SplitUp

People Meet

Embrace

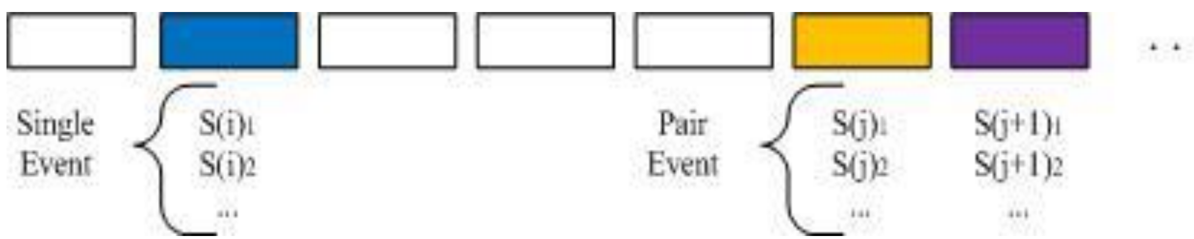
Pointing

Risk Ranking of Detected Events

- Objective:
  - To measure the risk of detections by considering: 1) the margin of top two classification candidates; 2) temporal relation; 3) potential gain of DCR;
  - Ranking data patterns by risk scores;
  - Checking and re-annotating the detections from high risk score to low risk score.

Risk Ranking of Detected Events

– Considering our classification results: for each segmentation S_i we have its top two candidates $\varphi^k(S_i)$ and $\varphi^{k'}(S_i)$, and their priors $p(k)$ and $p(k')$

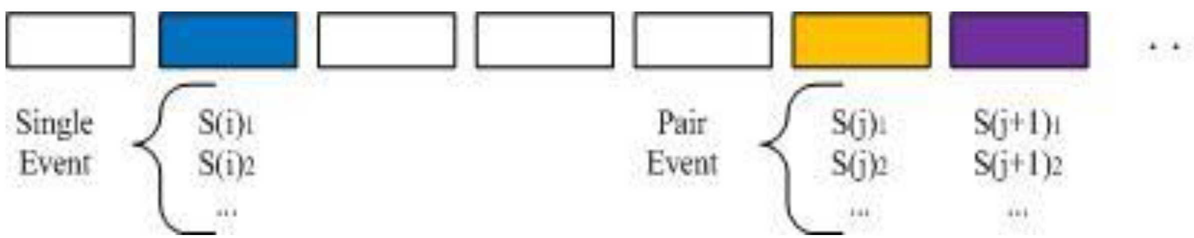


$$R(S_i) = \frac{1 - (\varphi^k(S_i)p(k) - \varphi^{k'}(S_i)p(k'))}{||S_i||} \cdot \begin{cases} w_m \\ w_f \\ w_m + w_f \end{cases}$$

w_m is the cost of a mis-detection and w_f is the cost of a false alarm, ($w_m = 1, w_f = 0.005$ were set based on DCR)

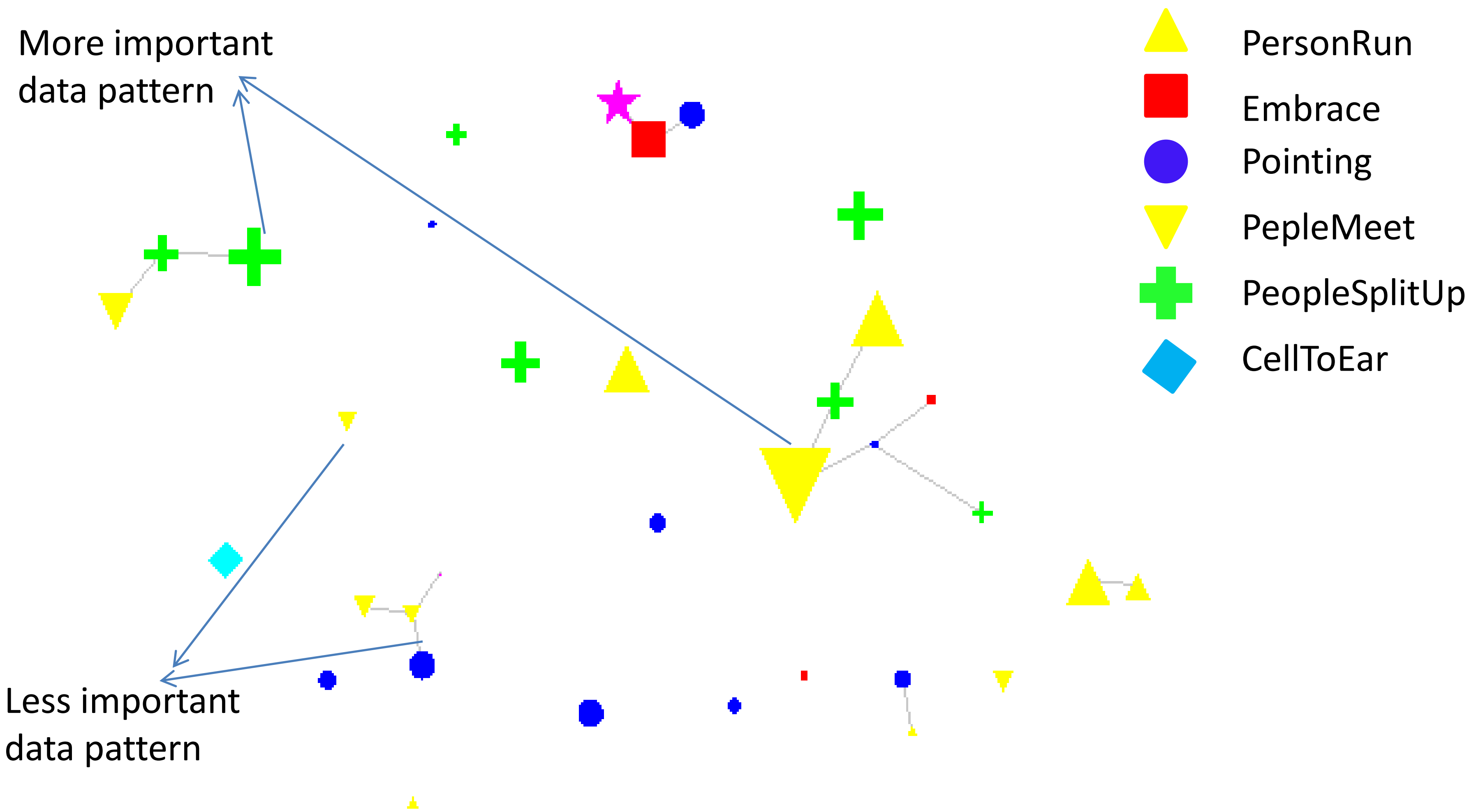
Risk Ranking of Detected Events

– Pair-wise events : for S_i and S_{i+1}, we have $\varphi^{k_j}(S_i)\varphi^{k_{j+1}}(S_{i+1})$ $\varphi^{k'_j}(S_i)\varphi^{k'_{j+1}}(S_{i+1})$ and their priors $p(k_j, k_{j+1})$ and $p(k'_j, k'_{j+1})$



$$R(S_i, S_{i+1}) = \frac{1-((\varphi^k(S_i)+\varphi^k(S_{i+1}))p(k_j,k_{j+1})-(\varphi^{k'}(S_i)+\varphi^k(S_{i+1}))p(k'_j,k'_{j+1})))}{\|S_i \cup S_{i+1}\|} \cdot \begin{cases} 2\cdot w_m \\ 2\cdot w_f \\ 2\cdot(w_m+w_f) \\ \ldots \end{cases}$$

Risk Ranking of Detected Events



Performance Evaluation

Actual DCR	Evaluation Set (25min * 7)			
	Retro	IBM-Inter-2014	IBM-Inter-2013	Others' Best 2014
CellToEar	0.9914	0.9849	0.9956	1.0013
Embrace	0.7456	0.6662	0.7337	0.6705
ObjectPut	1.0046	0.9960	0.9928	0.9705
PeopleMeet	0.8160	0.7965	0.9584	0.9094
PeopleSplitUp	0.8278	0.7869	0.8489	0.7918
PersonRuns	0.8111	0.8070	0.7188	0.6655
Pointing	1.0050	0.9788	0.9781	0.9725

- **Retro**: retrospective event detection system output.
- **IBM_Inter-2014**: primary run, risk ranking over all events, and interactive experiments are performed jointly with 175min .
- **IBM-Inter-2013**: performed separately for each event with 25 mins.
- **Others' Best 2014** :

Conclusions

- Retrospective System:
  - Joint-segmentation-classification provide a promising schema for surveillance event detection.
  - Modeling the long temporal relations can boost the detection performance.

- Interactive System:
  - Event visualization with strong temporal pattern can benefit the efficient interactive system.
  - Risk-based ranking of detected events with temporal pattern can boost the performance.

Future Works

- Retrospective System:
  - Exploiting deep learning for this task.
  - Exploring the performance trade-offs between localization and categorization.
- Interactive System:
  - Better visualization layout need to be developed, e.g. time layout.
  - Various risk ranking methods need to be tried.
  - User feedback utilization methods need to be incorporated. e.g. interactive learning.